Citizen Science Toolkit Conference

June 20 - 23, 2007

when pigs can really fly, we'll build a tool to count them

Steve Kelling
Director of Information Science
Cornell Lab of Ornithology

CORNELL LAB *of* ORNITHOLOGY

CORNELL LAB OF ORNITHOLOGY

607.254.BIRD telephone
www.birds.cornell.edu

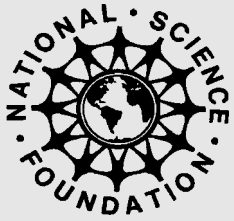
159 Sapsucker Woods Road
Ithaca, New York 14850

This presentation took place at the Citizen Science Toolkit Conference at the Cornell Lab of Ornithology in Ithaca, New York on June 20-23, 2007.

Note that this document did not originate as a formal paper. Rather, it combines an oral presentation with accompanying PowerPoint slides and reflects the more informal, idiosyncratic nature of a delivery prepared specifically for this live event.

Documentation of the conference is meant to serve as a resource for those who attended and for others in the field. It does not necessarily reflect the views of the Cornell Lab of Ornithology or individual symposium participants.

The following is the opening talk of the session titled "Technology and Cyberinfrastructure," delivered on day two of the Citizen Science Toolkit Conference



This documentation is supported by the **National Science Foundation** under Grant ESI-0610363.

Any opinions, findings, and conclusions or recommendations expressed in this documentation are those of the authors and do not necessarily reflect the views of the National Science Foundation.

For complete documentation of conference proceedings and to learn more about citizen science and the Citizen Science Toolkit, or to join the ongoing citizen science community, go to:

http://www.citizenscience.org

When Pigs Can Really Fly, We'll Build a Tool to Count Them

Steve Kelling,
Director of
Information Science,
Cornell Lab of Ornithology

The Impact of Advances in Technology

I really don't know much about counting pigs, but I bet I could figure out something. What I want to talk about are four related technologies that have really transformed how we manipulate information.

Advances in four related technologies have transformed how we access and manipulate information:

1. Moore's Law
2. The Internet and Web browsers
3. Global computer networks
4. Software

First, Moore's Law has made computers ubiquitous. The law predicts a doubling of the number of transistors on integrated circuits every eighteen months. This has led to rapid and continuing advances in computer power as well as lower unit costs.

The Internet and Web browsers have really created a global standard format in which we can pass information around between computers. Over the late '90s, advances in the amount of fiber optic cable that have been laid out across the world have made fiber optic cable a commodity that is allowing everyone to have access to computer networks. Finally, information description languages, data management processes, and software application integration have created seamless work flows for access, manipulation, and processing of almost limitless data resources.

What This Means for Citizen Science

What does all of this mean for citizen science? First, this has allowed us to create integrated project designs that allow us to record submissions, error checking, data management and data visualizations, all delivered through a single Web interface.

Second, this broad access and usability means that we can engage people in Panama or New Zealand just as easily as we can engage somebody in Peoria, Illinois. The scope of access is now global.

Third, data organization strategies have really allowed us to synthesize in markedly advanced ways that were unheard of several years ago.

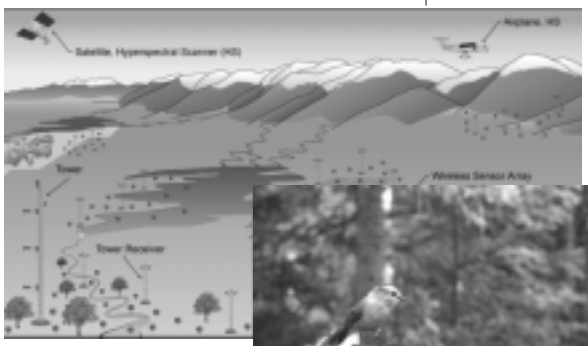
How do these advances impact participation in citizen science?

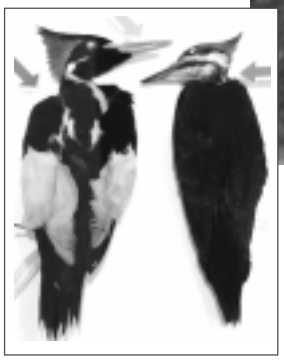
- Integrated project design
- Broad access and usability
- Data organization
- Dynamic information products

Finally, this all has allowed us to create a whole series of data information products, applications that allow a user to interact with all of this information in a number of ways and strategies that have never really occurred before.

The Focus: Observations

The way that I see it, the focus of citizen science is on observation. What we do is engage people in making observations of the world. What I mean by "observation" is a representation of the measurement of a defined attribute (e.g., a number, behavior, or physical property) of some "thing" observed. This could be an organism in the field, an observation of a bird flying overhead, a measurement of a specimen in a collection, or gathering of a sample during an experiment. The observation occurs at a particular time (e.g., date, time of day) and at a particular place (e.g., latitude and longitude of a point, transect, or polygon).





Observational data can be gathered using sensors or sensor arrays. It can be gathered by direct observations or by looking at biological specimen collections in museums. Observational data represent the greatest source of information on the distribution and abundance of organisms both across time and through space.

eBird

Origins and Overview

In 1997 the Lab, in collaboration with the National Audubon Society, initiated a project called BirdSource. The goal of the BirdSource project was to establish an ongoing data center capable of gathering, maintaining, analyzing, and broadly distributing the most current and comprehensive information on North American birds. While BirdSource has morphed into eBird, the basic goals of BirdSource are still the basic goals of the Information Science Program at the Lab.

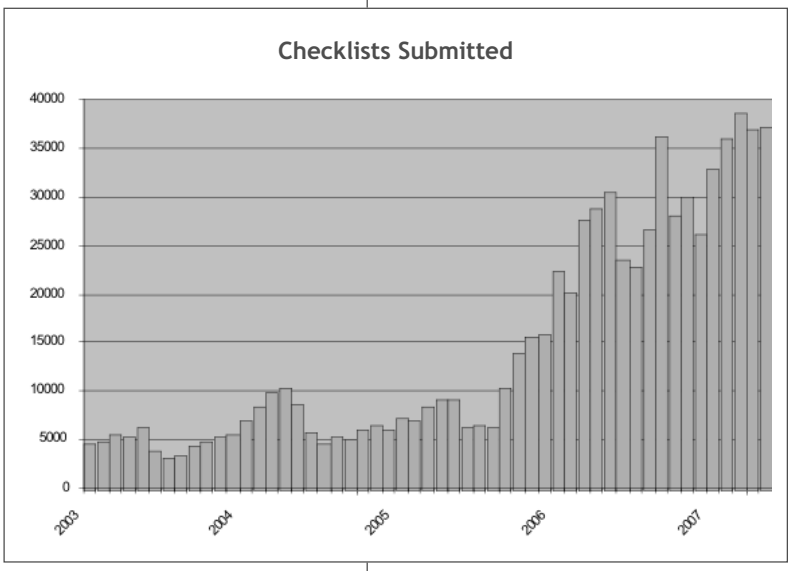
What I want to do today is describe the Lab's efforts in developing

this data center. Specifically, I want to talk about the enterprise application foundation of eBird, new directions we are heading in application development, and our efforts in the Avian Knowledge Network to organize and analyze observational data on bird populations.

What is eBird?

EBird is a hemisphere-wide checklist program whose goal is to engage birdwatchers to submit their observations in a standardized format. We began eBird in the fall of 2003 and are currently gathering over 35,000 checklists and 500,000 observations per month. Roughly about 10,000 individuals participate monthly.

EBird has been developed within an enterprise application framework. We use middleware software designed to set up, operate and integrate transaction-intensive applications across multiple computing platforms using Web technologies. We are using the same kinds of applications that banks use to process information to allow us to collect information via eBird.



Data are stored within an Oracle database management system, which allows us to create, update, and extract information very quickly, with much accuracy, and it can be scaled to be extremely accessible.

The most significant factor in eBird are its data gathering abilities, which have the potential of creating a network of bird observations at a global scale. However, if the data submission forms are not easy to use, then regardless of how sophisticated the data management or visualizations, not many people will participate and little information will be gathered.

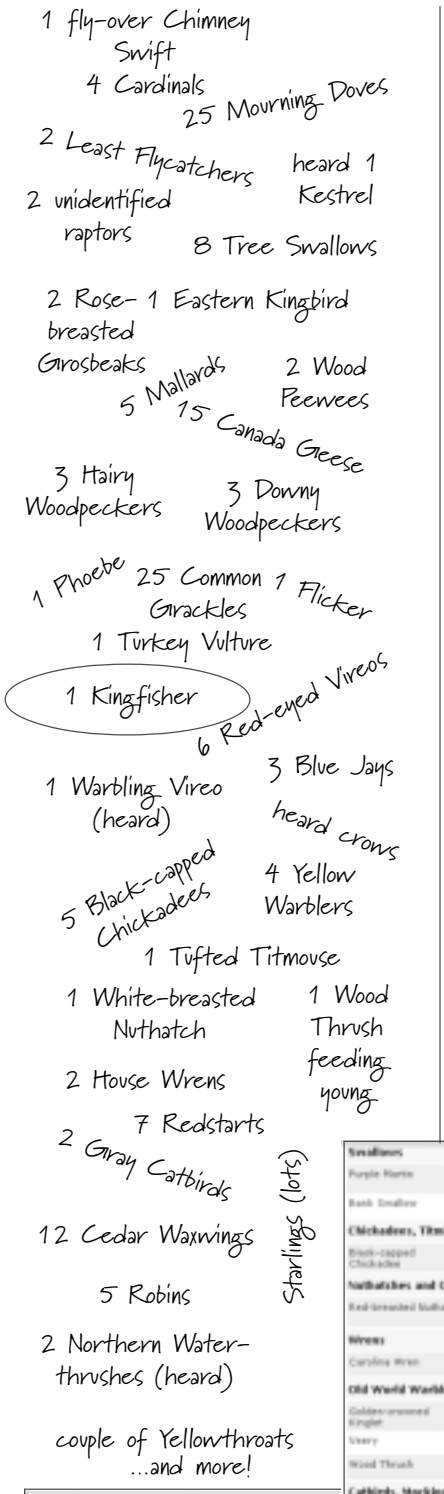
Sample Data Entry: A Conference Birdwalk

The process at eBird begins with an introductory page with lots of new features. You can submit observations, you can use a map or an

To demonstrate data entry, Kelling goes to the eBird Web site to enter observations from an early morning birdwalk by conference participants.

Try entering your own birding observations at:

http:www//eBird.org

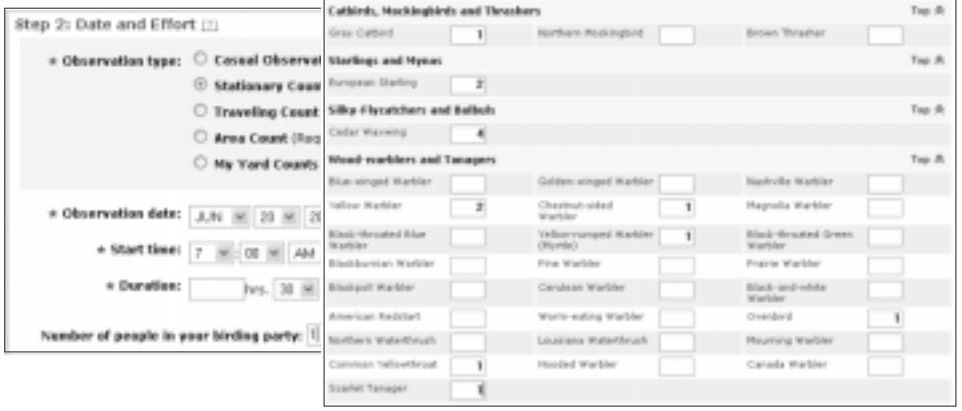


address, which we'll code, or you can go to specific locations where you've been birding before.

There are several protocols or observation types in eBird, and my guess is that the birding group that went out this morning did a traveling count—you walked around an area. All of this information including the location, the start time, the length of the walk, and number of people in the party allows us to standardize the way observations are gathered. One of the important pieces of information that we always ask for is whether you are submitting a complete checklist of the birds that you saw and heard. For example, you may have seen ten species of birds but didn't know what they were. Or you may have focused on just counting one species.

You then see a checklist. The checklists are specific for the region, in this case upstate New York, developed by an expert who knows the distribution and occurrence of birds across the year. This particular checklist is made for this area for the month of June. The checklists are being used by a range of age levels and abilities, from middle school students all the way up to expert users.

I've deliberately made an entry error and have recorded 175 Belted Kingfishers rather than 1. It flags our potential error and says, "175 is an impressive number...care to confirm?" We'll correct it, changing the 175 to a 1. Then we hit Continue and the list allows us to confirm what we saw. It asks, "Would you like to add a note?" We'll enter that this is a Citizen Science Toolkit expert bird team, and then we click Submit. We can e-mail this to the listserve. All of these data are now in the database so we can start looking at the results.



Maps and Checklists

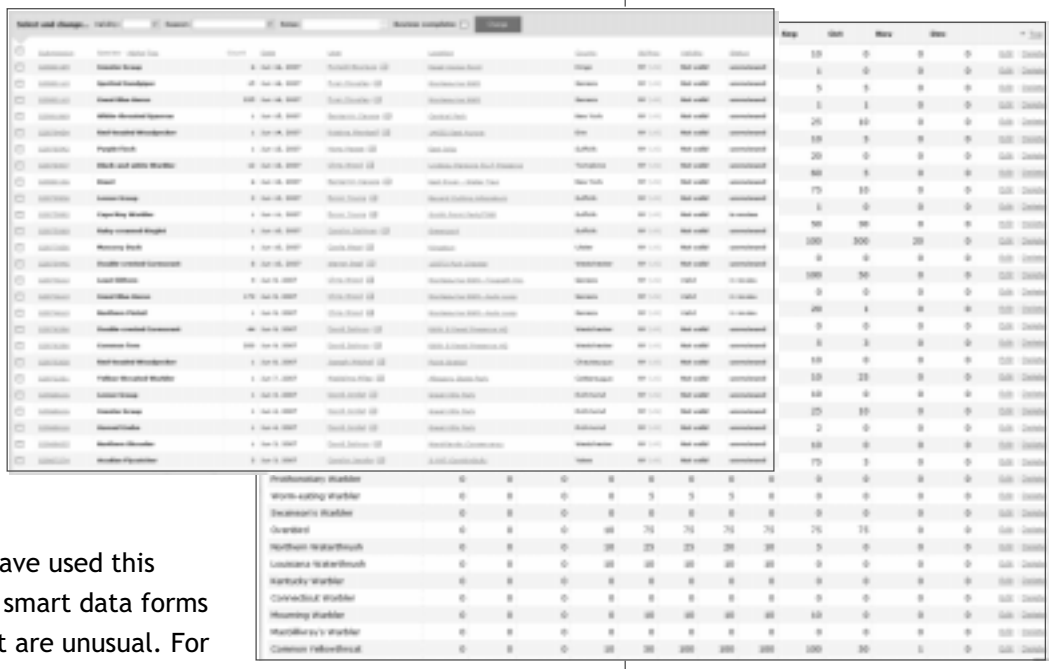
You saw during that demonstration examples of the kinds of filters we use. We probably have five or ten thousand of these filters created now for regions across the country, so the lists of birds we can present at any time are very accurate.

Users must be easily able to find the location in which they made their observations as accurately as possible. As I noted, this can be done

using a Google map, but it can also be done via known locations or address matching.

Data Quality and Expert Engagement

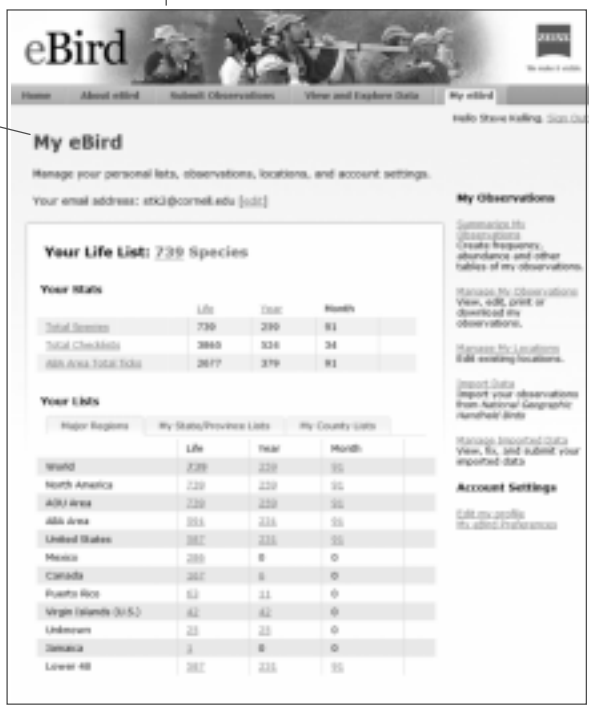
The other thing that we do is develop tools that allow us to engage regional experts to actually fill in the data, edit the data, and change the values of whether we accept the information that someone submits.



There is already much known about the distribution of bird populations in North America. We have used this knowledge to create smart data forms that flag records that are unusual. For many organisms (birds are an excellent example) expert opinion on species ranges and seasonal distributions are a valuable resource to improve data quality. Editing tools allow experts to view flagged records, contact the individual who made the observation, and validate or reject those records.

My eBird

Probably the most significant component that we've added that allowed us to increase the number of people participating is the "My eBird" feature. This is a way to allow anybody to keep track of the observations they've made, not only their life lists and things like that, but for particular locations. You can also generate graphs and maps of all of the observations that you've made.



Data Visualizations in eBird

The exciting thing about these kinds of occurrence data that we're collecting is that we can provide accurate estimations of the dynamics and patterns of abundance of bird populations at specific locations across North America. Because all eBird data are associated with a location, it is easy to organize

results in a variety of interactive visualizations. Histograms are generated dynamically from geo-referenced data.

Working with the New York Audubon Society, we have built an application that allows us to look at bird populations across all of the important New York bird areas. You can choose one or you can choose them all, and you can see how the patterns of things like the Least Sandpiper or Semipalmated Sandpiper appear in the spring as they migrate north and then spend some time at the Montezuma refuge in the fall until they leave.

Additionally, we are able to create maps showing the dynamics of bird populations across time and space. This is showing the pattern of a series of warblers in the winter, early spring, late spring, summer, early fall, late fall, and winter. [*Note: In original presentation, an animated version of the map below cycled through all seasons listed.*]

An interesting thing about this map is that because we ask that question, "Are you recording all the species you observed?" we also have information on where species weren't observed because we can infer that. All of the light gray squares that you see are locations where we have observations, but the species of interest weren't being reported. This is a significant advancement and a really relevant component of observational data that is important with the

types of data we are collecting with eBird.

The Need for Collaborative Development

The Enterprise System Dilemma

We have a problem, and I call this the "enterprise system dilemma." While the information architecture used for eBird has been successful, it does have its drawbacks. First, it is expensive to develop and maintain. Second, it takes a lot of effort to add or modify existing applications. Third, it is difficult to adapt this system to gather other types of data.

The Enterprise System Dilemma

- Observational data sets are extremely heterogenous
- Data standards vary across communities
- A need to improve data sharing and interoperability
- Develop a transparent application framework that is extensible and open source

Observational Data Systems

So we are starting, with some other groups, to look at how we can share the characteristics of observational data and develop systems that integrate these heterogeneous data sets and allow us to collect information across taxa. Additionally, data standards are currently varied across the observational community, making data integration and analysis difficult. We are interested in developing single data standards that will allow us to bring these kinds of data together.

Collaborative Development

These issues are not only being addressed by the Lab of Ornithology. NatureServe, which is managing most of the heritage programs of the Nature Conservancy and collecting that kind of data also have these kind of issues.

The National Center of Ecological Analysis and Synthesis (NCEAS), which works a lot with the long-term ecological research stations across the country also face these issues. These three groups are now getting together. We are going to have a meeting in July in Santa Barbara, and quite a few people from a broad range of backgrounds are going to be coming to begin to discuss and develop ways in which we can collaborate and bring these kinds of data together.

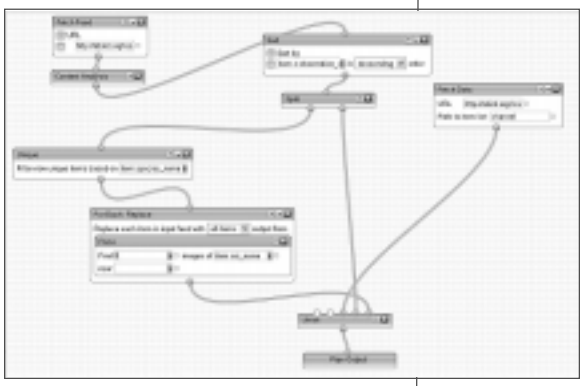
Our goal for this collaboration is to build a core observation data ontology. We want to make sure this ontology is extensible to allow flexibility in the different ways that data are gathered. That is going to allow us to develop an integrated project framework based on this core model.

Goals of the Collaboration

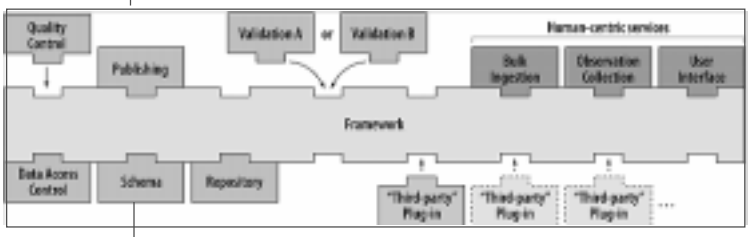
1. Develop a core observations data model
2. Allow the model to be extensible
3. Create an integrated project framework using the core model

Plugable Service Architecture

The concept that we are thinking about is development of a plugable service architecture to create a novel observation network. A user interface similar to Google Gadgets or Yahoo Pipes will provide an intuitive and straightforward environment that connects applications and code within the basic system framework. For example, you would have pieces of application code of something like the gadgets in Google Gadgets that a user, via a Web browser, can connect to and then create a kind of citizen science project or observational data gathering tool that they are interested in developing. They can then go online and start collecting these data.



We are going to have an application framework and, via this Web browser interface, an individual can add quality control features or different ways in which they want to collect, display and store these data. The concept is that this will all be built in an open source environment, so anybody can build additional gadgets or modify gadgets to increase the functionality. Anyone developing a data gathering application will be able to use these off-the-shelf services, extend existing services, or create entirely new services.

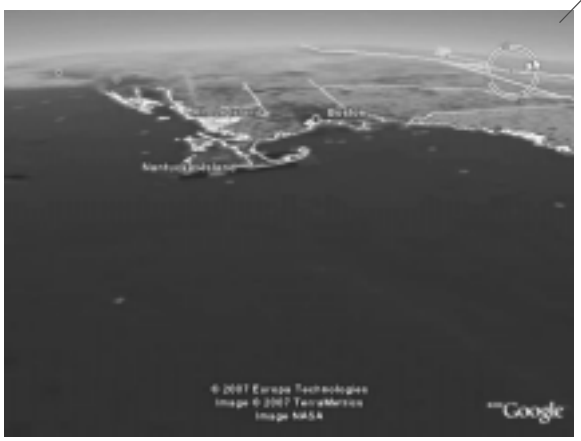


Lessons from Gathering Observational Data on Birds

Observational Data

We have already had great success in the bird monitoring community organizing and making available a rich and diverse data resource on bird occurrence. This is a fly-over of North America, and these red dots [*visible in animated fly-over*] are all the locations where we have information on the occurrence of bird populations.

North American Fly-over



This is a huge resource. Right now there are probably over 30,000 records that we have stored in the Avian Knowledge Network. We have records across broad geographic areas. It is sparse in some regions where there aren't many people, but we have a lot of data in Los Angeles. As you see during this fly-over, the concept is to bring all of these data into a single resource.

The Avian Knowledge Network

But we are really not interested in the data itself, we are interested in converting those kinds of data into knowledge, so we decided not to call this the Avian Data Network but instead the Avian Knowledge Network, primarily because we want to do more than just collect all of these data. The Avian Knowledge Network is an effort to convert these data into easily accessible knowledge for use in conservation and management of birds.

Mining for Conservation in the ~~Avian Data Network~~ Knowledge

We are motivated by the fact that all of these data, these counts of birds collected by various people and organizations, are potentially valuable beyond the data's initial use and potentially of even higher value when combined with similar data collected by others. This past week we just brought in ten years of breeding bird survey data. We just brought in some data from a watershed in Northern California. Our concept is that if we can integrate these data sources into a single repository that we could then analyze, the synergy of bringing those data together will allow us to explore the causes of species occurrence, bird populations, and current problems.

www. avianknowledge.net

This has allowed us to create a whole series of different kinds of exploratory analyses. This is a map of occurrence of high Arctic breeding shorebirds through the years across North America [*animated*].

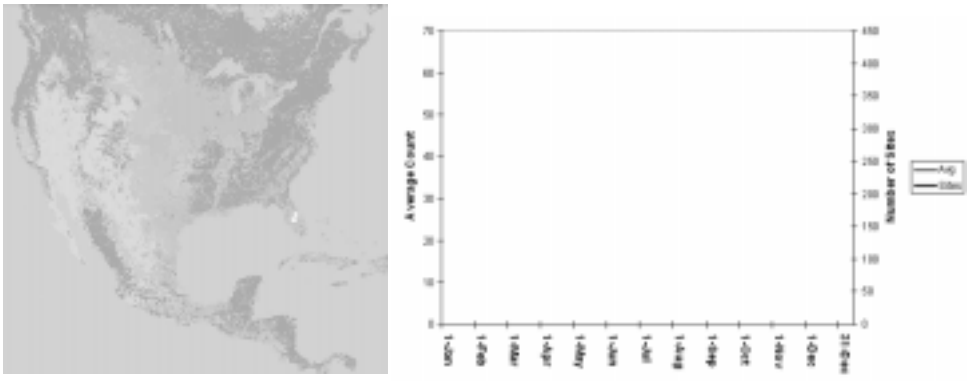
Migration of High Arctic Breeding Shorebirds

This was generated to begin modeling the spread of avian influenza by some researchers here. What you see by looking at these maps is

the occurrence, then the breeding up in the high Arctic, and then there is movement down across the continent.

Additionally, we can do things like look at the difference in population densities of Purple Martin across the United States, so during the breeding period you can see [*in animated version*] there are lots of locations where we have data on Purple Martin, but the red line shows that they are in low numbers. Then in July that changes. The average number reported increases rapidly, but the number of observations really drops.

Purple Martin



Exploratory Analysis

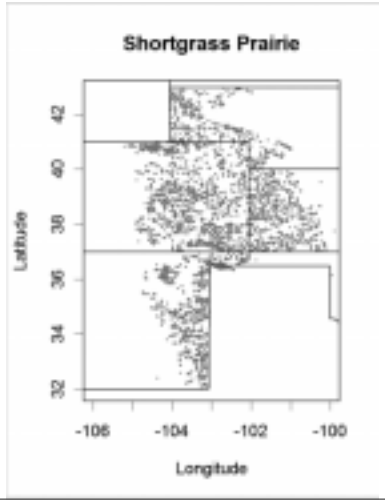
These kinds of applications, these new data mining tools that we're creating, allow us to look at the patterns of occurrence where we can relate precipitation of rainfall or cultivated crops—actually about 1,200 environmental variables. They are also allowing us to look at what influences bird populations.

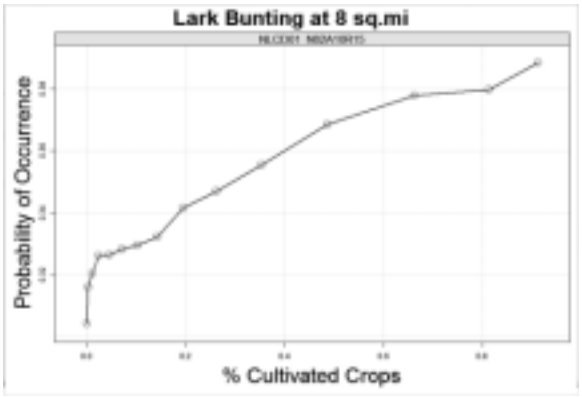
Use Data Mining models to generate environmental requirement profiles for bird populations across broad geographic regions.

30,000 observations
- Rocky Mountain Bird Observatory
- 2001-2005
- 10,000 locations

138 Predictors
- NLCD Habitat
  - 21 classes
  - 6 scales
- Climate (EPA)
  - Average Precipitation
  - Average Snowfall

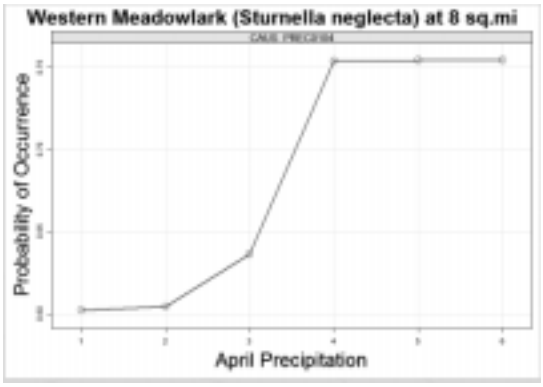
30 grassland species



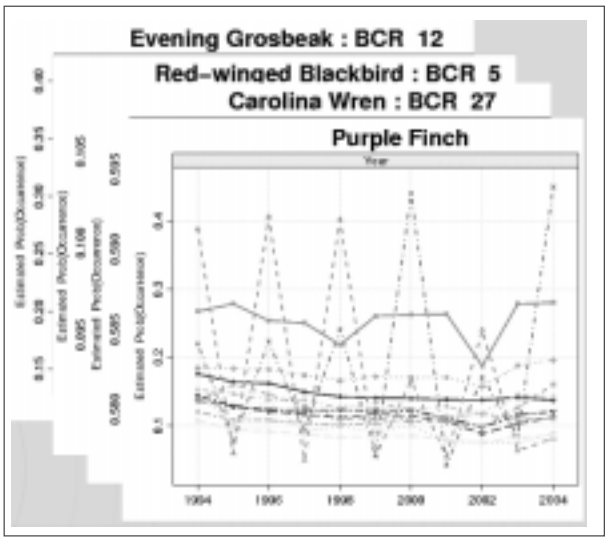


Lark Bunting
(Calamospiza melanocorys)

Western Meadowlark
(Sturnella neglecta)



The cool thing that we are doing is taking all of these data mining and analyses results and building a really simple user interface to a fairly sophisticated statistical package that will allow people to explore the patterns of species occurrence, to look at the impacts of population density for a year across bird populations.



Conclusions:

- Engage citizens to participate in observational data networks.
- Networks gather enormous volumes of data.
- The cyberinfrastructure is becoming more flexible and adaptive.
- The challenge is to convert these data into knowledge.
- Develop sound biodiversity conservation strategies.

Conclusion

In conclusion, what I really want to say is that we can now develop citizen-based sensor networks that can gather observational data at continental scales. These networks gather an enormous volume of data and the same engineering and systems management techniques used to calibrate and maintain vast autonomous sensor networks must be employed to insure the data quality.

The cyberinfrastructure that organizes these observational data networks is evolving and becoming more flexible and adaptive. The challenge now is integrating data from disparate resources and developing visualizations and analytical techniques that convert data into knowledge.

The ultimate goal is to develop sound biodiversity conservation strategies at all scales and for all organisms.

This is one of my favorite places. It is the view out my window here at the Lab.

All there is to thinking is seeing something noticeable, which makes you see something you were not noticing, which makes you see something that isn't even visible.
Norman Maclean

Every day I make observations of the birds and the weather and submit those observations to eBird. My goal is to develop these kinds of opportunities to allow anybody to submit their observations into these continental-scale database systems so that we can begin to explore and conserve biodiversity across the continent.